

統計学のポイント整理

野村友和

July 9, 2025

確率変数と確率分布

階乗

1 から n までの整数をかけ合わせたものを n の階乗といい， $n!$ で表す。ただし $0! = 1$ とする。

(例)

$$10! = 10 \times 9 \times 8 \times \cdots \times 1 = 3628800$$

順列

n 個の要素の中から k 個を取り出して順番に並べるとき，並べ方が何通りあるかを順列といい， ${}_nP_k$ で表す。

$${}_nP_k = \frac{n!}{(n-k)!}$$

(例) 5 個の要素の中から 3 個を取り出して並べるとき，並べ方は

$${}_5P_3 = \frac{5!}{(5-3)!} = 5 \times 4 \times 3 = 60 \text{通り}$$

組合せ

n 個の要素の中から k 個を取り出すとき、組合せが何通りあるかを ${}_nC_k$ で表す。

$${}_nC_k = \frac{{}_nP_k}{k!} = \frac{n!}{k!(n-k)!}$$

(例) 5 個の要素の中から 3 個をとり出すとき、組合せは

$${}_5C_3 = \frac{5!}{3!(5-3)!} = 10 \text{通り}$$

二項定理

$$\begin{aligned}(p+q)^n &= {}_nC_0 p^n q^0 + {}_nC_1 p^{n-1} q^1 + {}_nC_2 p^{n-2} q^2 + \cdots {}_nC_n p^0 q^n \\ &= \sum_{k=0}^n {}_nC_k p^{n-k} q^k\end{aligned}$$

確率変数と確率分布

どの値が実現するかはわからないが、実現しうるすべての値について、その確率がわかっている変数を**確率変数**という。

(例) サイコロを投げたとき、どの目が出るかはわからないが、1~6 までそれぞれの目が出る確率はわかっているので、サイコロを投げて出る目は確率変数。

確率変数に付された確率の系列を**確率分布**という。たとえば、サイコロを投げて出る目の確率分布は下のように表される。

事象	1	2	3	4	5	6
確率	1/6	1/6	1/6	1/6	1/6	1/6

確率変数は大文字（例： X ）で、その実現値は小文字（例： x ）で表記する。

離散型確率分布と連続型確率分布

サイコロの出目 X は 1~6 の 6 通りの値をとりうるが、1.5 や 2.5 などの値をとることはないので、起こりうる事象には番号をつけて数えることができる¹。このような確率分布を**離散型確率分布**という。

それに対して、たとえば生まれてくる赤ちゃんの体重は、2500g と 3000g の間の 2750g という値もとりうる。さらに、2750g と 2751g の間の 2750.5g という値もとりうるし...起こりうる事象は連続的で番号を付けて数えることができない。このような確率分布を**連続型確率分布**という。

確率分布が離散型であるか連続型であるかによって、確率の定義が異なる。

- 離散型確率分布では、確率変数がとりうる値のすべてにその値が実現する確率に対応し、**確率密度 = 確率**
- 連続型確率分布では、確率変数のとりうる値に確率ではなく確率密度に対応²し、**確率密度 $\neq$ 確率**

¹起こりうる事象の数は無限であってもよい（可付番無限）。

²連続型確率分布では、確率変数がとりうる値のすべてに正の確率を与えると確率の合計が無限大になってしまう。

離散型確率分布における確率・期待値・分散

$X = x$ という事象が生じる確率を、 $P(X = x)$ または $P(x)$ と書く。

確率分布の平均を**期待値**といい、確率変数 X の期待値を $E(X)$ と表す。また、確率変数 X の分散を $\text{Var}(X)$ と表す。

離散型確率分布の期待値：

$$E(X) = \sum_x xP(x)$$

確率変数の分散：

$$\text{Var}(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$$

(例) サイコロの出目 X の期待値と分散は

$$E(X) = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \cdots + 6 \times \frac{1}{6} = 3.5$$

$$\text{Var}(X) = (1 - 3.5)^2 \times \frac{1}{6} + (2 - 3.5)^2 \times \frac{1}{6} + \cdots + (6 - 3.5)^2 \times \frac{1}{6} = \frac{17.5}{6}$$

連続型確率分布における確率・期待値・分散

確率変数の値に確率密度を対応させる関数を密度関数といい、 $f(x)$ と書く。

連続型確率分布では、確率変数が a と b の間の値をとる確率が、密度関数の下側の面積で定められる。

$$P(a < X < b) = \int_a^b f(x)dx$$

したがって、連続型確率変数がある特定の値をとる確率 $P(x) = 0$ 。

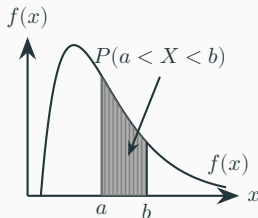


Figure 1: 連続型確率分布における確率

連続型確率変数の期待値：

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

連続型確率変数の分散：

$$\text{Var}(X) = E[(X - E(X))^2] = E(X^2) - [E(X)]^2$$

(例) 0～1 までの値をとる一様分布。

確率の合計は 1 となるので密度関数は

$$\begin{cases} f(x) = 1 & (0 < x < 1) \\ f(x) = 0 & (\text{それ以外}) \end{cases}$$

この一様分布の平均，分散は

$$E(X) = \int_0^1 x dx = \left[\frac{x^2}{2} \right]_0^1 = \frac{1}{2}$$

$$E(X^2) = \int_0^1 x^2 dx = \left[\frac{x^3}{3} \right]_0^1 = \frac{1}{3}$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = \frac{1}{3} - \left(\frac{1}{2} \right)^2 = \frac{1}{12}$$

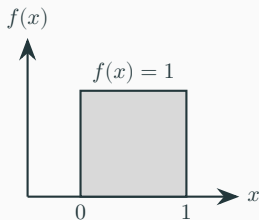


Figure 2: 一様分布の確率密度関数

積率母関数

確率変数 X について、 X^θ の期待値を θ 次の積率（モーメント）という。

次のような関数 $M(\theta)$ を離散型確率変数 X の積率母関数という。

$$M(\theta) = \sum_x e^{\theta x} P(x)$$

e^z をマクローリン展開すると

$$e^z = 1 + z + \frac{z^2}{2!} + \frac{z^3}{3!} + \frac{z^4}{4!} + \cdots$$

なので、積率母関数は以下のように書ける。

$$\begin{aligned} M(\theta) &= \sum_x \left(1 + \theta x + \frac{\theta^2 x^2}{2!} + \frac{\theta^3 x^3}{3!} + \frac{\theta^4 x^4}{4!} + \cdots \right) P(x) \\ &= \sum_x P(x) + \theta \sum_x x P(x) + \frac{\theta^2}{2!} \sum_x x^2 P(x) + \frac{\theta^3}{3!} \sum_x x^3 P(x) + \frac{\theta^4}{4!} \sum_x x^4 P(x) + \cdots \end{aligned}$$

積率母関数を θ について微分し, $\theta = 0$ とおくと

$$\left. \frac{dM(\theta)}{d\theta} \right|_{\theta=0} = \sum_x xP(x) = E(X)$$

さらに, 積率母関数を θ について 2 階微分し, $\theta = 0$ とおくと

$$\left. \frac{d^2M(\theta)}{d\theta^2} \right|_{\theta=0} = \sum_x x^2P(x) = E(X^2)$$

したがって, 積率母関数を θ について n 階微分し $\theta = 0$ とおくと, 確率変数 X の n 次のモーメントが得られる。

連続型確率分布の積率母関数は

$$M(\theta) = \int_{-\infty}^{\infty} e^{\theta x} f(x) dx$$

二項分布

1回の試行である事象が実現する確率を p とする。この試行を独立に n 回繰り返したとき、この事象が実現する回数 X が x 回となる確率は、

$$P(X = x) = {}_n C_x p^x (1 - p)^{n-x} = \frac{n!}{x!(n-x)!} p^x (1 - p)^{n-x}$$

このような離散型確率分布を**二項分布**といい、 $B(n, p)$ と書く。

(例) サイコロを 3 回投げて 1 の目が出る回数の分布は $B(3, \frac{1}{6})$

$$P(0) = {}_3 C_0 \times \left(\frac{1}{6}\right)^0 \times \left(\frac{5}{6}\right)^3 = \frac{125}{216}$$

$$P(1) = {}_3 C_1 \times \left(\frac{1}{6}\right)^1 \times \left(\frac{5}{6}\right)^2 = \frac{75}{216}$$

$$P(2) = {}_3 C_2 \times \left(\frac{1}{6}\right)^2 \times \left(\frac{5}{6}\right)^1 = \frac{15}{216}$$

$$P(3) = {}_3 C_3 \times \left(\frac{1}{6}\right)^3 \times \left(\frac{5}{6}\right)^0 = \frac{1}{216}$$

二項分布の積率母関数，期待値，分散：

$$\begin{aligned}M(\theta) &= \sum_{x=0}^n e^{\theta x} P(x) = \sum_{x=0}^n e^{\theta x} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \\&= \sum_{x=0}^n \frac{n!}{x!(n-x)!} (pe^{\theta})^x (1-p)^{n-x} \\&= (pe^{\theta} + (1-p))^n\end{aligned}$$

$$\begin{aligned}E(X) &= \left. \frac{dM(\theta)}{d\theta} \right|_{\theta=0} = n(pe^{\theta} + (1-p))^{n-1} pe^{\theta} \Big|_{\theta=0} \\&= np\end{aligned}$$

$$\begin{aligned}E(X^2) &= \left. \frac{d^2 M(\theta)}{d\theta^2} \right|_{\theta=0} = npe^{\theta} (pe^{\theta} + (1-p))^{n-1} \\&\quad + npe^{\theta} (n-1)(pe^{\theta} + (1-p))^{n-2} pe^{\theta} \Big|_{\theta=0} \\&= np + n(n-1)p^2\end{aligned}$$

$$\text{Var}(X) = E(X^2) - [E(X)]^2 = np(1-p)$$

正規分布

以下のような密度関数をもつ確率分布を**正規分布**という。

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right]$$

正規分布の特徴

- 左右対称のベル型（平均・中央値・最頻値が同じ）。
- 平均は μ ，分散は σ^2 。

確率変数 X が平均 μ ，分散 σ^2 の正規分布にしたがうとき，以下のように書く。

$$X \sim N(\mu, \sigma^2)$$



Figure 3: 正規分布の密度関数

平均 0, 分散 1 の正規分布 $N(0, 1)$ を標準正規分布という。

標準正規分布の密度関数：

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$$

確率変数 X が $N(\mu, \sigma^2)$ にしたがうとき, X の標準化変量 Z は標準正規分布にしたがう。

$$Z = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$P(Z > z) = \alpha$ となるような z を, z_α と書き $\alpha \times 100\%$ 点という。

標準正規分布は左右対称なので, $P(Z > z_\alpha) = P(Z < -z_\alpha) = \alpha$ である。

多変量の確率分布

二つ以上の確率変数が同時にとりうる値に対しても、一変数の確率分布と同様に確率または確率密度を対応させることができる。これを同時確率分布という。

たとえば、離散型確率変数 X が x という値をとり、かつ Y が y という値をとる確率を、 $P(X = x, Y = y)$ または $P(x, y)$ と書く。

X, Y の同時確率分布は以下の表のようにまとめることができる。

$X \backslash Y$	y_1	y_2	$\dots$	y_m	$P(x)$
x_1	p_{11}	p_{12}	$\dots$	p_{1m}	$p_{1\cdot}$
x_2	p_{21}	p_{22}		p_{2m}	$p_{2\cdot}$
$\dots$					$\dots$
x_n	p_{n1}	p_{n2}		p_{nm}	$p_{n\cdot}$
$P(y)$	$p_{\cdot 1}$	$p_{\cdot 2}$	$\dots$	$p_{\cdot m}$	1

Y がどの値をとるかに関わらず, X が x_i という値を取る確率 $P_X(X = x_i)$ を X の**周辺確率**という。

$$P_X(X = x_i) = \sum_{j=1}^m p_{ij} = p_{i.}$$

同様に Y の周辺確率は :

$$P_Y(Y = y_j) = \sum_{i=1}^n p_{ij} = p_{.j}$$

条件付き分布

$Y = y_j$ という条件の下での X の分布を, X の Y についての条件付き分布という。

$Y = y_j$ の下で X が x_i という値をとる確率を $P(X = x_i | Y = y_j)$ と表すと

$$P(X = x_i | Y = Y_j) = \frac{P(X = x_i, Y = y_j)}{P(Y = y_j)}$$

条件付き分布の期待値および分散を, 条件付き期待値および条件付き分散という。

$$E(X | Y = y_j) = \sum_{i=1}^n x_i \frac{P(x_i, y_j)}{P(y_j)}$$

$$\begin{aligned} \text{Var}(X | Y = y_j) &= E[(X - E(X))^2 | Y = y_j] \\ &= \sum_{i=1}^n [x_i - E(X | Y = y_j)]^2 \frac{P(x_i, y_j)}{P(y_j)} \end{aligned}$$

$$\text{または } \text{Var}(X | Y = y_j) = E(X^2 | Y = y_j) - [E(X | Y = y_j)]^2$$

以下が成立するとき， X と Y は統計的に独立であるという。

$$P(x_i, y_j) = P(x_i)P(y_j)$$

確率変数どうしが互いに独立であれば，各変数の条件付き分布が，ほかの確率変数がどのような値をとるかに依存しない（周辺分布と等しくなる）。

【定理】独立な確率変数の積

X ， Y が互いに独立であれば

$$E(XY) = E(X)E(Y)$$

証明：

$$\begin{aligned} E(XY) &= \sum_i \sum_j x_i y_j P(x_i, y_j) = \sum_i \sum_j x_i y_j P(x_i) P(y_j) \\ &= \sum_i x_i P(x_i) \sum_j y_j P(y_j) = E(X)E(Y) \end{aligned}$$

X と Y の共分散は以下で定義される。

$$\begin{aligned}\text{Cov}(X, Y) &= E[(X - E(X))(Y - E(Y))] \\ &= E(XY) - E(X)E(Y)\end{aligned}$$

X と Y の相関係数は以下で定義される。

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \cdot \text{Var}(Y)}}$$

【定理】独立な確率変数の共分散

X と Y が互いに独立であれば, $\text{Cov}(X, Y) = 0$

証明: X と Y が互いに独立であれば $E(XY) = E(X)E(Y)$ なので

$$\text{Cov}(X, Y) = E(XY) - E(X)E(Y) = 0$$

【定理】 確率変数の和の期待値

$$E(X + Y) = E(X) + E(Y)$$

【定理】 確率変数の和の分散

$\text{Cov}(X, Y) = 0$ であれば

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$$

確率変数の差の分散も $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y)$ となることに注意。

連続型の確率変数の同時確率分布では、離散型確率変数の場合の同時確率 $P(x, y)$ に同時確率密度 $f(x, y)$ が対応する。

連続型の同時確率分布の例：多変量正規分布

確率変数 X と Y が以下の同時確率密度関数を持つとき、 X と Y は二変量正規分布にしたがうという。

$$f(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho_{XY}^2}} \times \exp \left[-\frac{1}{2(1-\rho_{XY}^2)} \left(\frac{(x-\mu_X)^2}{\sigma_X^2} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} - \frac{2\rho_{XY}(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} \right) \right]$$

X の周辺確率密度関数を $f_X(x)$ と書くと

$$\begin{aligned} f_X(x) &\equiv \int_{-\infty}^{\infty} f(x, y) dy \\ &= \frac{1}{\sqrt{2\pi}\sigma_X} \exp \left[-\frac{1}{2} \left(\frac{x - \mu_X}{\sigma_X} \right)^2 \right] \end{aligned}$$

となり、平均 μ_X 、分散 σ_X^2 の正規分布となる。

X の期待値、分散はそれぞれ

$$\begin{aligned} E(X) &\equiv \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} x f_X(x) dx = \mu_X \end{aligned}$$

$$\begin{aligned} \text{Var}(X) &\equiv \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X))^2 f(x, y) dx dy \\ &= \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx = \sigma_X^2 \end{aligned}$$

同様に Y の期待値, 分散はそれぞれ

$$\begin{aligned} E(Y) &\equiv \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} y f(x, y) dy dx \\ &= \int_{-\infty}^{\infty} y f_Y(y) dy = \mu_y \end{aligned}$$

$$\begin{aligned} \text{Var}(Y) &\equiv \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - E(Y))^2 f(x, y) dy dx \\ &= \int_{-\infty}^{\infty} (y - \mu_Y)^2 f_Y(y) dy = \sigma_Y^2 \end{aligned}$$

X と Y の共分散は

$$\begin{aligned} \text{Cov}(X, Y) &\equiv \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - E(X))(y - E(Y)) f(x, y) dx dy \\ &= \rho_{XY} \sigma_X \sigma_Y \end{aligned}$$

X と Y の相関係数は

$$\text{Corr}(X, y) = \rho_{XY}$$

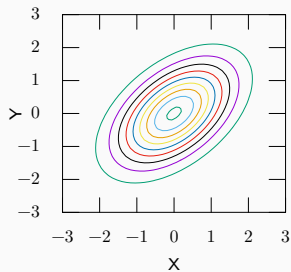
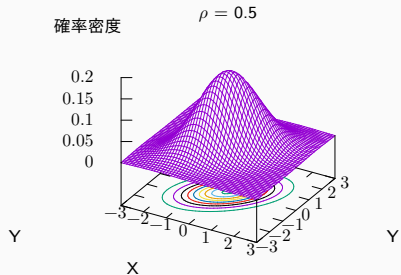
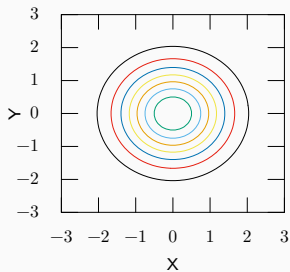
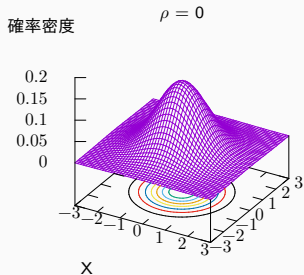


Figure 4: 二変数正規分布の密度関数と等高線

標本分布

推測統計は、母集団から抽出された標本を用いて母集団の性質を推測する。

われわれの関心は母集団であるが、通常は母集団全体を調査することはできない。

- 製品の耐久性の調査のように、調査が製品の破壊を意味する。
- 仮に調査が可能であっても、時間や費用がかかる。

そこで、母集団から標本を抽出して調査（標本調査）し、標本（母集団の一部）から母集団に関する情報を推測。

母集団の特性を表す値（母平均 μ や母分散 σ^2 など）のことを母数といい、標本にもとづく統計量（標本平均や標本分散など）を標本統計量³という。標本統計量は、標本抽出をやり直すたびに異なる値をとるので確率変数。標本統計量の分布を標本分布という。

³統計量は値を求めるための手続き（計算式）を指し、統計値は実際に計算された値を指す。

推定のために用いられる標本統計量を推定量という。

未知の母数を θ ，その推定量を $\hat{\theta}$ のように書く。

不偏性 : $\hat{\theta}$ が θ の不偏推定量であるとは，以下が成立することである。

$$E(\hat{\theta}) = \theta$$

一貫性 : $\hat{\theta}$ が θ の一致推定量であるとは，以下が成立することである。

$$\text{plim}_{n \rightarrow \infty} \hat{\theta} = \theta$$

有効性 : $\hat{\theta}$ が有効推定量であるとは，ほかのあらゆる推定量 $\tilde{\theta}$ に対して，以下が成立することである。

$$\text{Var}(\hat{\theta}) < \text{Var}(\tilde{\theta})$$

【定理】 標本平均の分布

$X_1, X_2, \dots, X_n$ を平均 μ , 分散 σ^2 の母集団からの大きさ n の無作為標本とする。このとき、標本平均 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ の期待値は μ , 分散は σ^2/n となる。

【定理】 正規母集団の標本平均の分布 (1)

$X_1, X_2, \dots, X_n$ を $N(\mu, \sigma^2)$ にしたがう母集団からの大きさ n の無作為標本とする。このとき、標本平均 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ は $N(\mu, \sigma^2/n)$ にしたがう。

中心極限定理

【定理】 中心極限定理

母集団がどのような分布であっても、 n 大きくしていくと、標本平均の分布は $N(\mu, \sigma^2/n)$ に近づいていく。

たとえば、サイコロを 1 回投げて出る目の分布は一様分布だが、サイコロを n 回投げて出る目の平均（標本平均）の分布は n が大きくなると

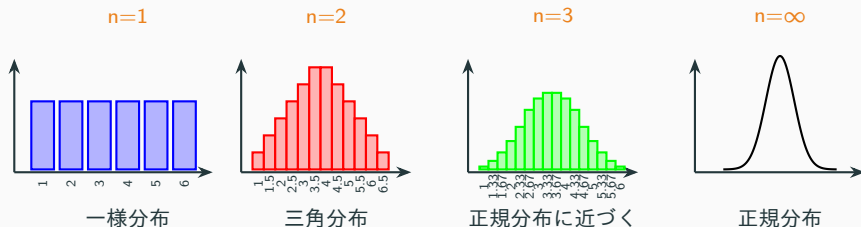


Figure 5: サイコロを投げる回数と出目の平均の分布

標本分散 S^2 を以下のように定義する。

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

【定理】 標本分散の期待値

$X_1, X_2, \dots, X_n$ を平均 μ , 分散 σ^2 の母集団からの大きさ n の無作為標本とする。このとき、標本分散 $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ の期待値は σ^2 となる。

すなわち、標本分散 S^2 は母分散 σ^2 の不偏推定量⁴。

⁴逆に、 $\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ は母分散を過小に推定。

証明：

$$\begin{aligned}\sum_{i=1}^n (X_i - \mu)^2 &= \sum_{i=1}^n [(X_i - \bar{X}) + (\bar{X} - \mu)]^2 \\ &= \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2 + 2(\bar{X} - \mu) \sum_{i=1}^n (X_i - \bar{X}) \\ &= (n-1)S^2 + n(\bar{X} - \mu)^2\end{aligned}$$

両辺に期待値をとると、 $E\left[\sum_{i=1}^n (X_i - \mu)\right] = n\sigma^2$ 、 $E\left[\sum_{i=1}^n (\bar{X} - \mu)\right] = \frac{\sigma^2}{n}$ より

$$n\sigma^2 = (n-1)E(S^2) + \sigma^2$$

よって

$$E(S^2) = \sigma^2$$

ただし、 $E(S) = \sigma$ とはならないことに注意⁵。

⁵ $f(X)$ が線型なら $E[f(X)] = f[E(X)]$ だが、 $f(X)$ が非線型の場合にはこれは一般に成立しない。

$Z_1, Z_2, \dots, Z_n$ が互いに独立に標準正規分布にしたがうとき

$$U = Z_1^2 + Z_2^2 + \dots + Z_n^2$$

は、自由度 n の χ^2 分布にしたがい、 $U \sim \chi^2(n)$ と書く。

したがって、 $X_1, X_2, \dots, X_n$ を $N(\mu, \sigma^2)$ にしたがう母集団からの無作為標本とするとき

$$\sum_{i=1}^n \left(\frac{X_i - \mu}{\sigma} \right)^2 \sim \chi^2(n)$$

さらに、以下の定理が成り立つ。

【定理】 S^2 を標本分散とすれば：

$$\frac{(n-1)S^2}{\sigma^2} = \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 \sim \chi^2(n-1)$$

Z が標準正規分布にしたがい、 U が自由度 k の χ^2 分布にしたがうとき、 Z と U が独立であれば、

$$T = \frac{Z}{\sqrt{U/k}}$$

は自由度 k の t 分布にしたがい、 $T \sim t(k)$ と書く。

【定理】正規母集団の標本平均の分布 (2) $X_1, X_2, \dots, X_n$ を平均 μ 、分散 σ^2 の母集団からの大きさ n の無作為標本とする。このとき、標本平均を $\bar{X}$ 、標本分散を S^2 とすれば

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t(n-1)$$

すなわち、母分散 σ^2 が未知の場合、標準化変量の σ を標本標準偏差 S で置き換えた統計量は、自由度 $n-1$ の t 分布にしたがう。

証明：

$$\begin{aligned} T &= \frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{S^2/\sigma^2}} \\ &= \frac{\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \end{aligned}$$

ここで、 $\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$ であるので、 T は自由度 $n-1$ の t 分布にしたがう。

t 分布は標準正規分布と同様の左右対称な分布であるが、自由度が小さいほど中心の密度が薄く、両裾の密度が厚い。そして、自由度が大きくなるにしたがって、標準正規分布に近づいていく。

U が自由度 m の χ^2 分布にしたがい、 V が自由度 n の χ^2 分布にしたがうとき、 U と V が独立であれば

$$F = \frac{U/m}{V/n}$$

は自由度 m, n の F 分布にしたがい、 $F \sim F(m, n)$ と書く。

【定理】正規母集団から大きさ n_1, n_2 の無作為標本を抽出する。このとき、それぞれの標本分散を S_1^2, S_2^2 とすれば

$$\frac{S_1^2}{S_2^2} \sim F(n_1 - 1, n_2 - 1)$$

推定

母数の区間推定と信頼区間

母数の**推定**：未知の母数を，標本から得られる情報で推測する。

たとえば，未知の母平均を標本から推測する場合，未知の母平均 μ を，ズバリ言い当てることはできないが，推定値に幅をもたせれば，一定の確率でその区間の中に母平均が入っている。

(例)

母平均を 10 と推定	→ 的中している確率はゼロ
母平均を 9.5～10.5 と推定	→ 的中している確率が 50%
母平均を 9～11 と推定	→ 的中している確率が 90%
母平均を 8～12 と推定	→ 的中している確率が 99%

的中する確率を高くしようと思えば，推定値の幅を広くしてやればよい。

確率 $1 - \alpha$ で的中するような，推定値の幅を確率 $1 - \alpha$ の**信頼区間**という。

信頼区間を求めることを**区間推定**という。

母平均の区間推定：正規母集団で母分散が既知の場合

正規分布にしたがう母集団から抽出されたサイズ n の標本の標本平均 $\bar{X}$ は、平均 μ 、分散 σ^2/n の正規分布にしたがうので、その標準化変量

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

は標準正規分布にしたがう。

仮に、母集団が正規分布で母分散がわかっているいれば⁶：

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

⁶未知の母集団の平均を推定するわけだから、母分散だけがわかっているというのはあまり現実的ではないが、ベンチマークとして考える。

書き直すと

$$P\left(\bar{X} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right) = 1 - \alpha$$

となるので、 $\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$ の区間に母平均 μ が含まれる確率が $1 - \alpha$ 。

標本平均 $\bar{X}$ の値が $\bar{x}$ であれば、母平均 μ の $(1 - \alpha) \times 100\%$ 信頼区間は

$$\left(\bar{x} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad , \quad \bar{x} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right)$$

標準正規分布表から、よく使う値を書き出しておく。

Table 4: 標準正規分布の臨界値

α	$z_{\alpha/2}$
0.01	2.576
0.05	1.960
0.10	1.645

母平均の区間推定：正規母集団で母分散が未知の場合

母集団の分散 σ^2 が未知の場合には、標本分散 S^2 を用いる。

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

σ を S で置き換えると

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

は自由度 $n-1$ の t 分布にしたがう。

したがって

$$P\left(-t_{\alpha/2}^{(n-1)} < \frac{\bar{X} - \mu}{S/\sqrt{n}} < t_{\alpha/2}^{(n-1)}\right) = 1 - \alpha$$

書き直すと

$$P\left(\bar{X} - t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}} < \mu < \bar{X} + t_{\alpha/2}^{(n-1)} \frac{S}{\sqrt{n}}\right) = 1 - \alpha$$

標本平均，標本分散の値がそれぞれ $\bar{x}$, s であれば，母平均 μ の 95% 信頼区間は

$$\left(\bar{x} - t_{\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}} \quad , \quad \bar{x} + t_{\alpha/2}^{(n-1)} \frac{s}{\sqrt{n}}\right)$$

t 分布は自由度が大きくなるにしたがって標準正規分布に近づいていくので，サンプル・サイズ n が十分に大きい場合には， $t_{\alpha/2}$ の代わりに $z_{\alpha/2}$ を用いてもよい。

母集団の分布が正規分布でなくても，サンプル・サイズ n が十分に大きければ標本平均の分布は正規分布で近似できるので， n が十分に大きければ，非正規母集団でも同じ方法で母平均の区間推定が可能。

X を 0 と 1 の二値をとる確率変数とし、母集団において $P(X = 1) = p$, $P(X = 0) = 1 - p \equiv q$ とする。

$X_1, X_2, \dots, X_n$ を母集団からの無作為標本とすれば

$$\sum_{i=1}^n X_i \sim B(n, p)$$

n が大きいとき、二項分布 $B(n, p)$ は正規分布 $N(np, npq)$ で近似できるため、標本比率 $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ の分布は、 $N(p, pq/n)$ で近似できる。

標準化すれば

$$Z = \frac{\bar{X} - p}{\sqrt{pq/n}} \sim N(0, 1)$$

したがって

$$P\left(-z_{\alpha/2} < \frac{\bar{X} - p}{\sqrt{pq/n}} < z_{\alpha/2}\right) = 1 - \alpha$$

書き直すと

$$P\left(\bar{X} - z_{\alpha/2}\sqrt{pq/n} < p < \bar{X} + z_{\alpha/2}\sqrt{pq/n}\right) = 1 - \alpha$$

ここで、 p, q は未知なので、 $\hat{p} = \bar{x}, \hat{q} = 1 - \bar{x}$ で置き換えれば近似的に

$$P\left(\bar{X} - z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n} < p < \bar{X} + z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}\right) = 1 - \alpha$$

母集団比率 p の $(1 - \alpha) \times 100\%$ 信頼区間は近似的に⁷

$$\left(\bar{x} - z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}, \quad \bar{x} + z_{\alpha/2}\sqrt{\hat{p}\hat{q}/n}\right)$$

⁷厳密には $(\bar{X} - p)^2 < z_{\alpha/2}^2 \cdot \frac{p(1-p)}{n}$ となるので、これを整理した

$$(n + z_{\alpha/2}^2)p^2 - (2n\bar{X} + z_{\alpha/2}^2)p + n\bar{X}^2 < 0$$

を満たすような p の範囲が信頼区間である。

(上式の不等号を等号に変えて p について解いて得られる二つの解が上限、下限となる)

$X_1, X_2, \dots, X_n$ を正規母集団 $N(\mu, \sigma^2)$ からの無作為標本とすると

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$$

であるから

$$P\left(\chi_{1-\alpha/2}^2(n-1) < \frac{(n-1)S^2}{\sigma^2} < \chi_{\alpha/2}^2(n-1)\right) = 1 - \alpha$$

書き直すと

$$P\left(\frac{(n-1)S^2}{\chi_{\alpha/2}^2(n-1)} < \sigma^2 < \frac{(n-1)S^2}{\chi_{1-\alpha/2}^2(n-1)}\right)$$

となり、母分散の $(1 - \alpha) \times 100\%$ 信頼区間が得られる⁸。

⁸ χ^2 分布は左右対称ではないので、上側と下側の両方の $\alpha/2\%$ 点が必要になることに注意。

仮説検定

母集団の性質に関する仮説が正しいかどうかを、標本から得られる情報に基づいて統計的に検証することを**仮説検定**という。

(例) ある蚊取り線香のメーカーは、自社製品の平均燃焼時間 μ は 6 時間であると主張している。しかし、実際に使ってみると、燃焼時間は 6 時間よりも短いような気がする。そこで、20 本の蚊取り線香の燃焼時間を標本調査したところ、標本平均は 5 時間 45 分であった。このとき、メーカーの主張が正しいのかを統計的に検証しよう。

メーカーの主張： $\mu = 6.0$

われわれの主張： $\mu < 6.0$

メーカーの主張 ($\mu = 6.0$) が正しいもとでは、確率変数 $\bar{X}$ が 5.75 という値をとる確率が十分に小さいことを示せば、メーカーの主張 (仮説) を棄却することができる。

棄却したい仮説のことを**帰無仮説**といい、帰無仮説を棄却することにより主張したい仮説を**対立仮説**という。

Step 1 : 帰無仮説 (H_0) ・ 対立仮説 (H_1) を設定する。

今の例では

$$H_0 : \mu = 6.0$$

$$H_1 : \mu < 6.0$$

帰無仮説を**棄却**することができれば、平均燃焼時間が6時間よりも短いという対立仮説を**採択**する。

帰無仮説を棄却することができなければ、帰無仮説を受容する（平均燃焼時間が6時間であるという可能性を否定できないだけであって、積極的に帰無仮説が正しいと主張するわけではない）。

Step 2 : 有意水準を決める。

仮説検定には，二種類の過誤が生じる。

- 第一種の過誤：帰無仮説が正しいにもかかわらず，これを棄却してしまう。
- 第二種の過誤：帰無仮説が正しくないにもかかわらず，これを受容してしまう。

第一種の過誤が生じる確率 α を有意水準という。

- できるだけ過誤が生じないようにしたければ α を小さく設定すればよい（帰無仮説は棄却されにくくなる）。
- α を大きくすれば帰無仮説は棄却されやすくなるが，過誤が生じる確率が高くなる。

有意水準はコンテキストに応じて決めればよいが，0.05（5%），0.10（10%），0.01（1%）がよく用いられる。

Step 3 : 検定統計量の値を求める。

蚊取り線香の燃焼時間の母集団分布は正規分布で、分散が 0.5^2 であることがわかって
いるとしよう。

このとき、もし帰無仮説が正しければ、標本平均 $\bar{X}$ の分布は $N(6, 0.5^2/20)$ なので、
標準化変量 $Z = \frac{\bar{X}-6}{0.5/\sqrt{20}}$ は標準正規分布にしたがう（この Z を検定統計量という）。

今、標本平均 $\bar{X}$ の実現値 $\bar{x} = 5.75$ であるから、検定統計量 Z の実現値 z は

$$z = \frac{5.75 - 6}{0.5/\sqrt{20}} = -2.236$$

Step 4. 棄却域を求め、帰無仮説を棄却することができるかを判断する。

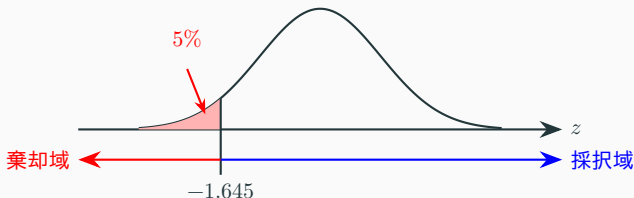


Figure 6: 棄却域

$P(Z < -1.645) = 0.05$ なので、帰無仮説が正しいもとで、 Z が -1.645 よりも小さくなる確率は 5%。

検定統計量の値 z が、臨界値 -1.645 よりも小さければ、5% の確率でしか生じない事象が起きたというよりは、帰無仮説が正しくないと考えの方が妥当。

→ z が臨界値 -1.645 よりも小さければ、帰無仮説を棄却する。

棄却域：帰無仮説を棄却する検定統計量の範囲

採択域：帰無仮説を採択する検定統計量の範囲

今の例では、 $z = -2.236$ は棄却域に入るので、蚊取り線香の平均燃焼時間が 6 時間であるという帰無仮説は棄却される。

対立仮説と棄却域

棄却域は，対立仮説に応じて決定する。

$H_0 : \mu = \mu_o$ に対して

- $H_1 : \mu < \mu_o$ なら棄却域は左側
- $H_1 : \mu > \mu_o$ なら棄却域は右側
- $H_1 : \mu \neq \mu_o$ なら棄却域は両側

注意：対立仮説や有意水準はあらかじめ決めておくべきもので，検定統計値を見てから決めるのではない。

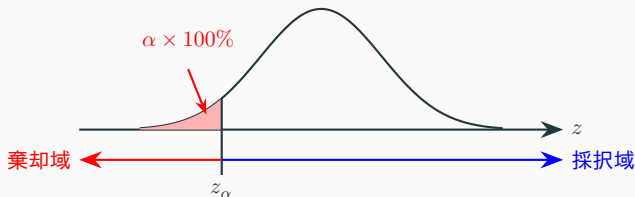


Figure 7: $H_1 : \mu < \mu_o$ の場合の棄却域

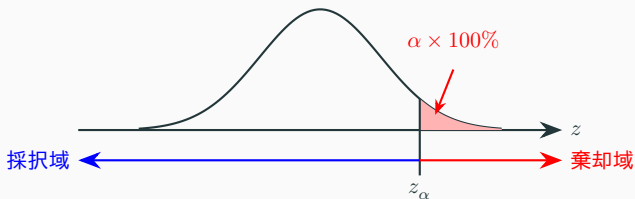


Figure 8: $H_1 : \mu > \mu_o$ の場合の棄却域

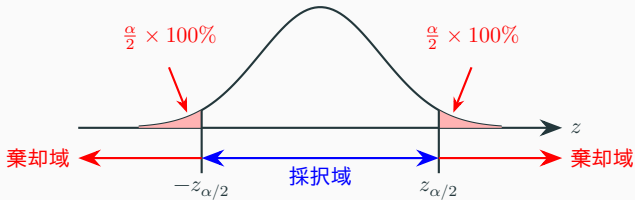


Figure 9: $H_1 : \mu \neq \mu_o$ の場合の棄却域

平均値の検定：正規母集団で母分散が未知の場合

正規母集団で母分散 σ^2 が未知の場合には、標本分散を用いて検定を行う。

先述の例で蚊取り線香の燃焼時間の分布は正規分布だが、分散は未知であるとする。

帰無仮説と対立仮説は先ほどと同様

$$H_0 : \mu = 6.0$$

$$H_1 : \mu < 6.0$$

20 本の標本調査により、 $\bar{x} = 5.75$, $s^2 = 0.75^2$ であったとしよう。

検定統計量は $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$ であり、帰無仮説が正しいもとで自由度 $n - 1$ の t 分布にしたがう。

$t_{0.05}^{19} = 1.729$ であるから、検定統計量の値が -1.729 よりも小さければ、有意水準 5% で帰無仮説を棄却する。

$$t = \frac{5.75 - 6}{0.75/\sqrt{20}} = -1.49$$

検定統計量は棄却域に入らないので、帰無仮説を棄却することはできない。すなわち、平均燃焼時間が 6 時間よりも短いとはいえない。

t 分布は自由度が大きくなるにしたがって標準正規分布に近づいていくため、 n が十分に大きい場合には、標準正規分布の臨界値を用いてもよい。

二つの正規母集団 $N(\mu_1, \sigma_1^2)$ と $N(\mu_2, \sigma_2^2)$ の平均 μ_1 と μ_2 とが等しいかどうかを、それぞれの母集団から抽出された標本により検定する。

帰無仮説と対立仮説は

$$H_0 : \mu_1 = \mu_2$$

$$H_1 : (1)\mu_1 > \mu_2, \quad (2)\mu_2 > \mu_1, \quad (3)\mu_1 \neq \mu_2 \quad \text{のいずれか。}$$

それぞれの母集団から n_1, n_2 の標本を抽出し、標本平均を $\bar{X}_1, \bar{X}_2$ とする。

母分散 σ_1^2, σ_2^2 がいずれも既知であるとき、標本平均 $\bar{X}_1, \bar{X}_2$ の分布は

$$\bar{X}_1 \sim N(\mu_1, \sigma_1^2/n)$$

$$\bar{X}_2 \sim N(\mu_2, \sigma_2^2/n)$$

確率変数 $\bar{X}_1$, $\bar{X}_2$ は独立なので

$$\begin{aligned}E(\bar{X}_1 - \bar{X}_2) &= E(\bar{X}_1) - E(\bar{X}_2) = \mu_1 - \mu_2 \\ \text{Var}(\bar{X}_1 - \bar{X}_2) &= \text{Var}(\bar{X}_1) + \text{Var}(\bar{X}_2) = \sigma_1^2 + \sigma_2^2\end{aligned}$$

また、正規分布の再生性⁹により

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$$

標準化すると

$$\frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1)$$

帰無仮説が正しいもとでは、 $\mu_1 - \mu_2 = 0$ なので、検定統計量は

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} \sim N(0, 1)$$

⁹ 正規分布にしたがう変数の和および差はまた正規分布にしたがう。

有意水準 α の棄却域は，対立仮説に応じて

$$(1) Z > z_{\alpha}, \quad (2) Z < -z_{\alpha}, \quad (3) |Z| > z_{\alpha/2}$$

母分散 σ_1^2, σ_2^2 が未知の場合は，標本分散 S_1^2, S_2^2 で置き換えた検定統計量

$$Z = \frac{(\bar{X}_1 - \bar{X}_2)}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

が漸近的（ n_1, n_2 が十分に大きければ）に標準正規分布にしたがうことを利用して検定を行う。

比率の差の検定

二つの母集団における比率 p_1, p_2 が等しいかどうかを、それぞれの母集団から抽出された標本により検定する。

帰無仮説と対立仮説は

$$H_0 : p_1 = p_2$$

$$H_1 : (1)p_1 > p_2, \quad (2)p_2 > p_1, \quad (3)p_1 \neq p_2 \quad \text{のいずれか。}$$

それぞれの母集団から n_1, n_2 の標本を抽出し、標本比率を $\bar{X}_1, \bar{X}_2$ とする。

- $\bar{X}_1$ の平均は p_1 , 分散は $p_1(1-p_1)/n_1$
- $\bar{X}_2$ の平均は p_2 , 分散は $p_2(1-p_2)/n_2$

したがって、標本比率の差 $(\hat{p}_1 - \hat{p}_2)$ の平均および分散は

$$E(\bar{X}_1 - \bar{X}_2) = p_1 - p_2$$

$$\text{Var}(\bar{X}_1 - \bar{X}_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

ここで、 n_1, n_2 が十分に大きければ、 $\bar{X}_1, \bar{X}_2$ の分布はそれぞれ正規分布で近似できるので、 $(\hat{p}_1 - \hat{p}_2)$ も近似的に正規分布にしたがう。

帰無仮説のもとでの $p_1 = p_2$ の値を p とすれば、検定統計量は

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{p(1-p) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} \sim N(0, 1)$$

ただし、ここで p は未知なので、 $\hat{p} = \frac{\sum_{i=1}^{n_1} X_{1i} + \sum_{i=1}^{n_2} X_{2i}}{n_1 + n_2}$ で近似する。

有意水準 α の棄却域は、対立仮説に応じて

$$(1) Z > z_\alpha, \quad (2) Z < -z_\alpha, \quad (3) |Z| > z_{\alpha/2}$$

二つの正規母集団 $N(\mu_1, \sigma_1^2)$ と $N(\mu_2, \sigma_2^2)$ の分散 σ_1^2 と σ_2^2 が等しいかどうかを、それぞれの母集団から抽出された標本により検定する。

帰無仮説と対立仮説は

$$H_0 : \sigma_1 = \sigma_2$$

$$H_1 : \sigma_1 > \sigma_2$$

それぞれの母集団から n_1, n_2 の標本を抽出し、標本平均を $\bar{X}_1, \bar{X}_2$ とする。

このとき

$$\sum_{i=1}^{n_1} \left(\frac{X_{1i} - \bar{X}_1}{\sigma_1} \right)^2$$
$$\sum_{i=1}^{n_2} \left(\frac{X_{2i} - \bar{X}_2}{\sigma_2} \right)^2$$

はそれぞれ、自由度 n_1, n_2 の χ^2 分布にしたがう。

したがって、これらを自由度で割ったものの比は自由度 n_1, n_2 の F 分布にしたがう。

$$F = \frac{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(\frac{X_{1i} - \bar{X}_1}{\sigma_1} \right)^2}{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} \left(\frac{X_{2i} - \bar{X}_2}{\sigma_2} \right)^2} \sim F(n_1, n_2)$$

帰無仮説が正しいもとで、この統計量は

$$F = \frac{\frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (X_{1i} - \bar{X}_1)^2}{\frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (X_{2i} - \bar{X}_2)^2}$$

となり、 $F > F_{\alpha}^{(n_1-1, n_2-1)}$ であれば帰無仮説を棄却する。